A Simple Escape from Moral Twin Earth*

Pekka Väyrynen
University of Leeds

1. Introduction

Naturalist moral realism (NMR) says that some moral claims are true, true moral claims are made so by objective, stance-independent moral facts, and these moral facts fall into the class of natural facts. A moral realist must offer a *metasemantics* for moral language: a theory that predicts to what properties moral terms will refer in various possible scenarios. Many versions of NMR deploy a causal theory of reference to explain how moral terms like 'good' and 'right' get to refer to certain natural properties. The essence of the Moral Twin Earth (MTE) objection to NMR is that although causal theories of reference are plausible for at least proper names and natural kind terms, they're not a plausible metasemantics for moral terms (Horgan and Timmons 1991, 1992a-b).[1] Many philosophers think that the MTE objection makes any version of NMR which it touches hopeless. Two standard moves on behalf of NMR are to (a) defuse the intuitions that drive the MTE objection or (b) develop a semantics and/or pragmatics of moral language which aims to avoid the problem the MTE objection raises against NMR.[2] In this paper I'll first outline the MTE objection and then present a simple reply: the causal theory of reference for moral terms due to Boyd (1988), a central original target of the MTE objection, is in fact not vulnerable to the MTE objection.

* This is a pre-print version of a paper forthcoming in *Thought*. Thanks to three anonymous reviewers for helpful comments on earlier versions of this paper.

1 Horgan and Timmons (2009) extend the MTE objection to an alternative metasemantics for NMR, moral functionalism (e.g. Jackson 1998). I won't take a view on whether anything in this paper helps NMR avoid the MTE objection when combined with metasemantic views other than the causal theory.

2 The literature is vast. For (a), see e.g. Laurence et al. (1999), Merli (2002), and Dowell (2016); for (b), see e.g. Sayre-McCord (1997), Copp (2000), Brink (2001), and van Roojen (2006). McPherson (2013) provides a useful overview.

2. The Moral Twin Earth Objection

According to a causal theory of reference for a given term, the term has its reference determined by what its use is causally linked to in the appropriate way. (This way doesn't require speakers to associate any identifying description with the term). When appropriate causal links are in place, our uses of the term are "causally regulated" by the property that's the referent. Putnam (1975) asks us to imagine a Twin Earth that's identical to Earth except that the stuff in their lakes, taps, and bodies isn't H_2O but a chemical, XYZ, whose surface qualities are very similar to H_2O but whose micro-structure is radically different. Putnam argues that although we and our twins have nearly identical dispositions to apply our respective English and Twinglish words 'water' to clear potable liquids in our environments, the strong intuition is that these uses of 'water' refer to different properties: H_2O in our case, XYZ in theirs. This is well explained if the reference of 'water', as used by a linguistic community, is determined by what substance causally regulates their use of the term. Since that substance is H_2O in our case but XYZ on Twin Earth, 'water' refers to different substances on Earth and Twin Earth.[3]

Horgan and Timmons argue that a Moral Twin Earth thought experiment which closely parallels Putnam's Twin Earth thought experiment speaks against the view that the reference of moral terms is similarly determined by what their use is appropriately causally linked to.[4] The denizens of Moral Twin Earth have a vocabulary that works very much like moral vocabulary on Earth: those who speak Twin English use the terms 'good', 'bad', 'right', and 'wrong' to evaluate actions, persons, institutions, and so on. Twin Earthlings' use of these terms also bears the marks that are often invoked to characterize moral vocabulary and practice: our twins use them to reason about considerations bearing on well-being, are normally disposed

3 A bit more precisely, 'water' (as used by us) refers to whatever has an internal structure identical to that of the liquid samples initially dubbed 'water' – namely, anything that's H_2O – provided that communicative exchanges by which the speakers at a dubbing lend reference to others generate appropriate causal links between the initial H_2O samples and our use of 'water'.

4 Horgan and Timmons single out Boyd's theory in particular in their (1991: 453) and (1992a: 158).

to act in ways corresponding to judgments about what's 'good' and 'right', normally regard what's 'good' and 'right' as particularly important in deciding what to do, and so on (Horgan and Timmons 1991: 459). Given that these terms play the same practical role as the orthographically identical English terms play in the moral practices of Earthlings who speak English, Earthling visitors to Moral Twin Earth should be strongly inclined to treat them as moral terms. However, such visitors would also notice significant differences regarding what sentences involving terms like 'good' are affirmed by Twin Earthlings. Earthlings are (let's imagine) disposed to apply 'good' to options which maximize overall happiness, disregard options which clearly bring about less overall happiness than some alternative, and resent other agents for taking options that fail to maximize overall happiness. By contrast, Twin Earthlings are disposed to apply 'good' to options they can will as a universal law, but disregard options they cannot will as a universal law and resent others for taking such acts.

The MTE objection supposes that given these differences, the two communities' uses of 'good' are causally regulated by different properties: a consequentialist property like maximizing net happiness for the consequentialist community C on Earth, a deontological property like passing the categorical imperative test for the deontological community D on Moral Twin Earth. The causal theory then predicts that 'good' has different referents in C and D despite playing the same role in regulating deliberation and sentiments. Horgan and Timmons claim that in cases like this we have a clear intuition that 'good' has the same reference in both communities. If I claim that a given act (killing an innocent person to save many lives, perhaps) is 'good', and my twin claims it is 'not good', my twin is denying an ascription of the same property that

I'm ascribing.[5] This difference with the Twin Earth intuition about 'water' suggests that moral terms don't have the kind of metasemantics that Putnam offers for natural kind terms.

The general form of the MTE objection is thus as follows. We describe two linguistic communities such that the causal theory of reference predicts that a moral term 'T^C' used by linguistic community C (above, 'good' on Earth) and a moral term 'T^D' used by linguistic community D ('good' on Moral Twin Earth) refer to different properties even when the usage of these terms is otherwise as similar as possible. We then elicit a semantic intuition that if one person accepts that something is 'T^C' and another denies that it's 'T^D', they aren't referring to different properties.[6] Since the causal theory conflicts with the semantic intuition, we should reject the causal theory, and along with it any form of NMR that relies on it.

3. Why Boyd's Theory Escapes the MTE Objection

The escape pod for Boyd's theory of reference is an epistemic condition on reference which is typically not adequately registered in discussions of the MTE objection. Here is Boyd:

> *Roughly*, and for nondegenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k* (excuse the blurring of the use-mention distinction). (Boyd 1988: 195; cf. Boyd 2003: 515)

5 The claim that Horgan and Timmons make is that C and D disagree substantively, not merely in the way that I and my twin disagree if I say of a liquid in a dirty puddle that it's 'not water' and my twin calls a similar puddle 'water'. The status of this intuition has been discussed extensively; see McPherson (2013) for references. Also note that it's controversial that genuine disagreement requires co-reference (Plunkett and Sundell 2013). If it doesn't, the semantic intuition required by the MTE objection is more questionable.

6 The MTE objection denies the possibility that an utterance of 'x is good' in C is true iff x has the consequentialist property while an utterance of 'x is good' in D is true iff x has the deontological property. The semantic options this rules out aren't limited to the Putnam-style conclusion that the meaning of 'good' may be totally different in C and D. It also cannot be that the utterances in C and D are true relative to different perspectives, or that they have different context-sensitive semantic values.

When such causal links obtain between our use of ‘good’ and moral goodness, our use is regulated by moral goodness in a way that enables our beliefs increasingly to approximate truths regarding it as a result of subsequent moral and nonmoral reasoning. This tendency may fail to be manifested when our beliefs are distorted by biases or other interfering factors. But, degenerate cases aside, the causal links that are relevant to the reference of ‘good’ must provide sufficient epistemic access to moral goodness to form the basis for the growth of knowledge about what is morally good (Boyd 1988: 201). In less jargon-laden terms: roughly, ‘good’ refers to moral goodness just in case there are at least some people who, under ordinary circumstances, are at least pretty good at finding out about moral goodness and reflect this capacity in what they say using ‘good’. If this kind of epistemic condition isn’t met, our uses of ‘good’ won’t count as referring to moral goodness.

I’ll argue as follows. Communities C and D must either converge in their subsequent uses of ‘good’ or not. If they converge, no conflict arises with the semantic intuition Horgan and Timmons elicit in MTE scenarios. But if they don’t converge, it’s no longer intuitively clear that C and D refer to the same property. To begin, suppose that the consequentialist and deontological properties after which C’s and D’s uses of ‘good’ pattern aren’t coextensive. Now consider two scenarios:

> *Convergence:* C’s subsequent moral reasoning takes its use of ‘good’ closer to the deontological property, or D’s subsequent moral reasoning takes its use closer to the consequentialist property, or the subsequent moral reasoning in each converges toward a third property, G.

> *Divergence:* The subsequent moral reasoning in C and D doesn’t manifest convergence.

In the relevant scenarios neither convergence nor divergence occurs through chance or fluke. Divergence, for instance, is supposed to occur even in the absence of the sort of distorting influences on moral beliefs which may block a tendency to increasingly approximate moral truths, such as self-interest and various other biases. Boyd suggests that, given the subsequent convergence in moral beliefs, various historical divergences in moral views (regarding the moral status of slavery, the divine right of kings, and more) are plausibly interpreted as Convergence scenarios where the earlier divergences reflect the influence of such distorting factors (Boyd 1988: 209-14; cf. Railton 1986: 195-200).

The epistemic condition on reference doesn't preclude the possibility of even serious errors about which properties determine how we in practice classify things under a term.[7] What Boyd's restriction to "nondegenerate" cases does require is that the initial background beliefs of members of C and D are relevantly approximately true and their methods of reasoning are approximately reliable, since otherwise the tendency for their subsequent reasoning increasingly to approximate truths is undermined (cf. Boyd 1988: 189-91). So in the relevant Convergence scenarios, C and D both increasingly converge toward truths regarding moral goodness. Such a tendency leaves room for various sorts of differences. Two communities that have competing beliefs or theories about the same property can co-refer using 'good'. Two communities can also have disputes over competing methodologies for inquiring into the nature of the same property.[8]

The MTE objection has little force under Convergence. On Boyd's theory, if C and D increasingly approximate moral truths in non-flukey ways, this is very likely because their uses of 'good' are causally regulated by the same property and their differences merely reflect some competing views about the nature of that property. But then NMR agrees that 'good' has the

7 Boyd claims that Newtonians were talking about mass and energy all along despite being massively wrong about the nature of space-time (1988: 210-11).

8 Even a community that's going wrong but can come to see the need for reform using their existing methods of reasoning might be using 'good' in reasoning in a way that forms a basis for the growth of moral knowledge.

same reference in C and D. Under Divergence, however, each community's use of 'good' can increasingly approximate truths about moral goodness only up to a limit. Boyd's theory allows that C and D may refer to different properties if they would continue to differ significantly in their applications of 'good' even without the influence of distorting factors (cf. Horgan and Timmons 1992b: 243). This is supposed to be the wrong result. So the MTE objection seems more forceful in Divergence scenarios.

Moral realism implies that in Divergence at most one of C and D is on the right track to truths regarding moral goodness. According to Boyd's theory, this means that at most one of C and D has its use of 'good' causally regulated by *moral goodness*. The problem for the MTE objection is that if C and D don't end up approximating truths regarding moral goodness roughly equally closely, then the following symmetries between C and D cannot all hold, given the epistemic condition on reference: (i) 'good' plays the same role in their practices; (ii) their initial beliefs about what's 'good' approximate truths about moral goodness well enough not to compromise the subsequent growth of knowledge regarding moral goodness; and (iii) they're using roughly equally reliable methods of reasoning about what's morally good equally properly.[9] (i)-(iii) are core elements of Convergence scenarios. So we should expect at least one of them to fail in Divergence scenarios. But a nested trilemma shows that if one but not the other community's subsequent reasoning about what's 'good' fails to approximate truths regarding moral goodness through the failure of any of (i)-(iii), it isn't intuitively clear that 'good' has the same reference in C and D. But the MTE objection requires this to be intuitively clear.

The MTE objection founders if Divergence occurs through failure of (i). The intuition that C and D differ in their beliefs about what's 'good' but not in the reference of 'good' is heavily based on the stipulation that 'good' plays the same practical role for each community.

9 Horgan and Timmons (1992a: 165) stipulate that MTE scenarios involve equally proper and thorough uses of the *same* reliable methods of moral inquiry. The formulation in the text is weaker.

Nor can Divergence occur through failure of (ii). Putnam's original Twin Earth scenario suggests that if the Twin property were completely unfamiliar to us and satisfied no serious competing theory here on Earth about the nature of goodness, the intuition that 'good' has the same reference in C and D would be undermined.[10] But a significant difference between the Twin Earth and Moral Twin Earth scenarios is that while Earthlings have no causal contact with XYZ and Twin Earthlings have no causal contact with H_2O, members of C and D have causal contact with both the deontological and the consequentialist property. Many actions done in each instantiate (or else fail to instantiate) both properties. So we can expect each community's initial 'good' beliefs to overlap significantly, irrespective of which property causally regulates their uses of 'good'. Scenarios where each community's initial beliefs provide sufficient epistemic access to goodness to form the basis for the subsequent growth of moral knowledge are candidates for Convergence. So if Divergence occurs through failure of (ii), one of C and D must have more initial 'good' beliefs which are false of moral goodness than the other – sufficiently many to lack the tendency to approximate truths regarding moral goodness through subsequent reasoning about what's 'good'. But then it's no longer intuitively clear that C and D refer to the same property. We're supposed to imagine that: the two communities use equally reliable methods of reasoning about what's 'good' from a partially shared stock of approximately true beliefs regarding what's 'good'; one of them subsequently goes increasingly wrong regarding moral goodness in their beliefs (but not through fluke or bias); and yet 'good' has the same reference in C and D. It's unclear what we're asked to imagine. Intuitions about such scenarios shouldn't be trusted.

Nor can Divergence occur through failure of (iii). Here we're supposed to imagine that: each community's initial beliefs about what's 'good' approximate moral truths closely enough to

10 No one would say that the Twinglish word 'water' refers to XYZ and not H_2O if Putnam's Twin Earth had both H_2O and XYZ on it and H_2O played the roles it does on Earth (Laurence, Margolis, and Dawson 1999: 163).

provide a basis for the growth of knowledge regarding moral goodness; one community's subsequent reasoning about what's 'good' involves methods that aren't reliable with respect to moral goodness (but not through bias or fluke); and yet 'good' co-refers in C and D. The community that fails increasingly to approximate truths regarding moral goodness despite having causal contact with whatever property in fact constitutes moral goodness won't merely have an inferior theory of that property; that's compatible with Convergence. But if we're to imagine that the two communities don't use even roughly equally reliable methods for reasoning about moral goodness, it isn't intuitively clear that their uses of 'good' co-refer. It isn't clear that their 'good' discourse is aptly interpreted as involving reasoning about matters such as how we can effectively care about each others' well-being or flourishing in a socially rational way, or that 'good' otherwise plays the same role. They might be using methods that are inappropriate for reasoning about what's morally good. Or they might be using the same method as their twins (such as reflective equilibrium, perhaps) to inquire reliably into some property other than moral goodness (quoodness, perhaps).[11] But then approximations to truth in their reasoning would be approximations regarding some property other than moral goodness. So again it isn't intuitively clear that 'good' has the same reference in both communities, given what Divergence scenarios must be like under Boyd's theory of reference.

4. A Third Option?

One might object that Boyd's epistemic condition on reference is compatible with a third option.[12] In MTE scenarios the consequentialist and the deontological property overlap in a wide range of cases and 'good' plays the same role in the practices of C and D. But now suppose that some acts which instantiate both the consequentialist and the deontological property aren't

11 Thanks to a referee for *Thought* for the 'quoodness' example.
12 Thanks to a referee for *Analysis* for this objection.

called ‘good’ by either community. C and D might then meet the epistemic condition by each coming to apply ‘good’ accurately to these overlapping cases over time. Since they’re improving in the same ways, their uses will be no less similar than they were at the start. And yet their uses of ‘good’ might not converge: outside the extensional overlap, C might apply ‘good’ to C-but-not-D actions (those that instantiate the consequentialist but not the deontological property) and D might apply it to D-but-not-C actions. Boyd’s theory allows that ‘good’ has a different reference in the two communities, but intuitively the reference is supposed to be the same.[13] So the epistemic condition won’t rule out all MTE scenarios that are problematic for NMR.

This scenario isn’t relevantly distinct, however. At least one of these communities will fail to approximate certain truths regarding moral goodness, given their different beliefs about things that are C-but-not-D and D-but-not-C. For this to be a non-Convergence scenario, something must make at least one community’s use of ‘good’ fail to be regulated by moral goodness despite both having causal contact with the property that in fact constitutes moral goodness and applying ‘good’ in some ways that approximate truths regarding it. But now we have a Divergence scenario, and so can re-run the arguments above.

A natural worry concerns the interpretation of the epistemic condition. Boyd says that for a community’s use of ‘good’ to refer to moral goodness, they must approximate moral truths in such a way that “it is possible to see how continued approximations would be forthcoming as a result of subsequent moral and nonmoral reasoning” (Boyd 1988: 201). Reading this as requiring continued approximations to a complete set of truths about what things are morally good might be objectionably strong, given that continued approximation is compatible with having many false beliefs. But the reply at hand doesn’t require a strong reading like this. Consider things that are both C and D. Members of D will say that these things wouldn’t be

13 The case is complex enough that this is in fact not clear. If C’s and D’s moral views developed toward incommensurable ways of achieving goods like human flourishing, it might be appropriate to treat ‘good’ as partially denoting each of the two different versions of the good (Boyd 2003: 511, 547).

good if they were C-but-not-D, and vice versa for members of C. Such dependence conditionals are first-order moral claims. (Compare: 'Kicking dogs wouldn't be bad if it didn't cause pain and suffering' and 'Kicking dogs wouldn't be bad if we approved of it' are first-order moral claims.) So there will be many truths regarding moral goodness which at least one community is bound to fail to approximate even with respect to things that in fact are C and D, even given a promising start in their beliefs about what's 'good'. But differences in these kinds of moral beliefs are most naturally thought of as reflecting competing theories of the same property.

5. Conclusion

Standard formulations of the MTE objection against NMR target the causal theory of reference for moral terms. I've argued that NMR can avoid the MTE objection once we take proper account of an epistemic condition which was part of Boyd's causal metasemantics all along. The plausibility of an epistemic condition on reference isn't something I can settle here. But the condition doesn't seem unmotivated. Boyd motivates the condition for at least theoretical terms (among which he counts moral terms) by noting that if the relations that are relevant to reference allow us increasingly to approximate the truth, this helps to explain how our uses of theoretical terms, and the associated classificatory practices, contribute to the inductive and explanatory success of our theories. It also helps to explain why the fact that two linguistic communities apply different definitions or descriptive characterizations in using a term doesn't by itself show that they're referring to different things (Boyd 1988: 195). A related idea is a fundamental connection between reference and epistemic justification: if a body of justified beliefs is about something, this guarantees that when the beliefs fail to match their object, the situation is somehow unlucky (Dickie 2016). An epistemic condition on reference would follow naturally.

If my argument is good, the MTE objection marks no advance over older semantic arguments against NMR which appeal to the notion of referential stability. Moral terms are referentially stable if two terms that play the same conceptual role (such as, perhaps, regulating agents' deliberation and sentiments in a certain way) are thereby guaranteed to have the same reference.[14] One such older argument is R. M. Hare's argument against descriptivism based on his example of the missionary and the cannibals (Hare 1952: 148). We needn't invoke Moral Twin Earth to ask whether NMR is committed to denying the referential stability of moral terms. Nor do we need to invoke it to assess whether denying their referential stability would be problematic.[15] The related question of whether the practical role of moral terms could be explained in some other way if NMR had to deny the referential stability of moral terms is also of broader significance. If referential stability can be resisted, then so can metaethical theories which imply it, such as many forms of expressivism and conceptual role semantics. It would thus be valuable to integrate discussions of what sort of metasemantics should go with NMR into the broader metaethical context beyond the MTE objection. A more systematic treatment of the adequacy conditions for the metasemantics of moral language would be welcome in itself and might raise challenges that are less easy for moral realists to escape.

14 Cf. Williams (2018) on the referential stability of moral concepts (rather than terms).
15 This issue is discussed at length in Eklund (2017).

References

Boyd, R. “How to Be a Moral Realist,” in *Essays on Moral Realism*, edited by G. Sayre-McCord. Ithaca, NY: Cornell University Press, 1988, 181-228.

Boyd, R. “Finite Beings, Finite Goods: The Semantics, Metaphysics and Ethics of Naturalist Consequentialism.” *Philosophy and Phenomenological Research* 66 (2003): 505-53.

Brink, D. “Realism, Naturalism, and Moral Semantics.” *Social Philosophy and Policy* 18 (2001): 154-75.

Copp, D. “Milk, Honey, and the Good Life on Moral Twin Earth.” *Synthese* 124 (2000): 113-37.

Dickie, I. “The Essential Connection Between Epistemology and the Theory of Reference.” *Philosophical Issues* 26 (2016): 99-129.

Dowell, J. L. “The Metaethical Insignificance of Moral Twin Earth.” *Oxford Studies in Metaethics* 11 (2016): 1-27.

Eklund, M. *Choosing Normative Concepts.* Oxford: Oxford University Press, 2017.

Hare, R. M. *The Language of Morals.* Oxford: Clarendon Press, 1952.

Horgan, T. and Timmons, M. “New Wave Moral Realism Meets Moral Twin Earth.” *Journal of Philosophical Research* 16 (1991): 447-65.

Horgan, T. and Timmons, M. “Troubles for New Wave Moral Semantics: The ‘Open Question Argument’ Revived.” *Philosophical Papers* 21 (1992a): 153-75.

Horgan, T. and Timmons, M. “Troubles on Moral Twin Earth: Moral Queerness Revived.” *Synthese* 92 (1992b): 221-60.

Horgan, T. and Timmons, M. “Analytic Moral Functionalism Meets Moral Twin Earth,” in *Minds, Ethics, and Conditionals*, edited by I. Ravenscroft. Oxford: Oxford University Press, 2009, 221-36.

Jackson, F. *From Metaphysics to Ethics.* Oxford: Clarendon Press, 1998.

Kripke, S. *Naming and Necessity.* Oxford: Blackwell, 1980.

Laurence, S., Margolis, E., and Dawson, A. "Moral Realism and Twin Earth." *Facta Philosophica* 1 (1999): 135-65.

McPherson, T. "Semantic Challenges to Normative Realism." *Philosophy Compass* 8 (2013): 126-36.

Merli, D. "Return to Moral Twin Earth." *Canadian Journal of Philosophy* 32 (2002): 207-40.

Plunkett, D. and Sundell, T. "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13.23 (2013): 1-37.

Putnam, H. "The Meaning of 'Meaning'." *Midwest Studies in the Philosophy of Science* 7 (1975): 131–93.

Railton, P. "Moral Realism." *Philosophical Review* 95 (1986): 163-207.

Sayre-McCord, G. "'Good' on Moral Twin Earth." *Philosophical Issues* 8 (1997): 267-92.

van Roojen, M. "Knowing Enough to Disagree: A New Reply to the Moral Twin-Earth Argument." *Oxford Studies in Metaethics* 1 (2006): 161–194.

Williams, J. R. G. "Normative Reference Magnets." *Philosophical Review* 127 (2018): 41-71.